

# Data Mining and Knowledge Discovery

Petra Kralj Novak

[Petra.Kralj.Novak@ijs.si](mailto:Petra.Kralj.Novak@ijs.si)

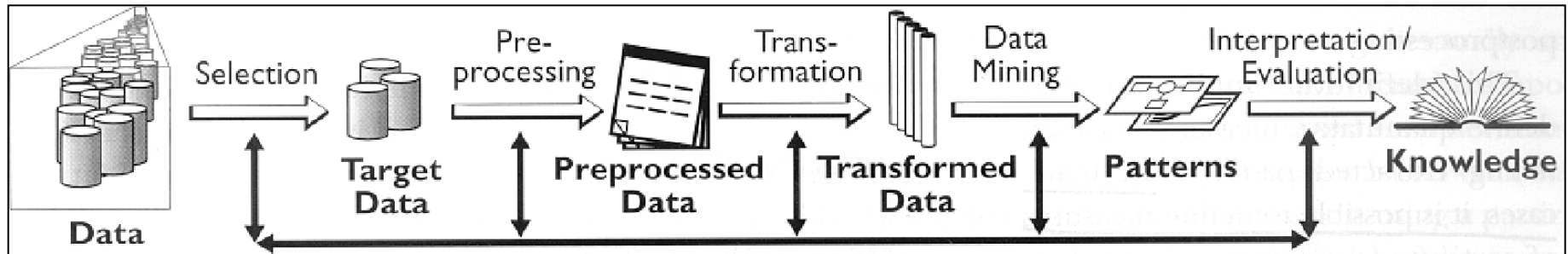
2011/11/22

# Practice plan

---

- 2011/11/08: Predictive data mining 1
  - Decision trees
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - A taste of Weka
- 2011/11/22: Predictive data mining 2
  - Evaluating classifiers 2: Cross validation
  - Naïve Bayes classifier
  - Numeric prediction
- 2011/11/29: Descriptive data mining
  - Association rules
  - Descriptive data mining in Weka
  - Discussion about seminars and exam
- 2011/12/20: Written exam, Seminar proposal presentations
- 2012/1/24 : Data mining seminar presentations

# Keywords



- Data

- Attribute, example, target variable, class, train set, test set, attribute-value data, market basket data

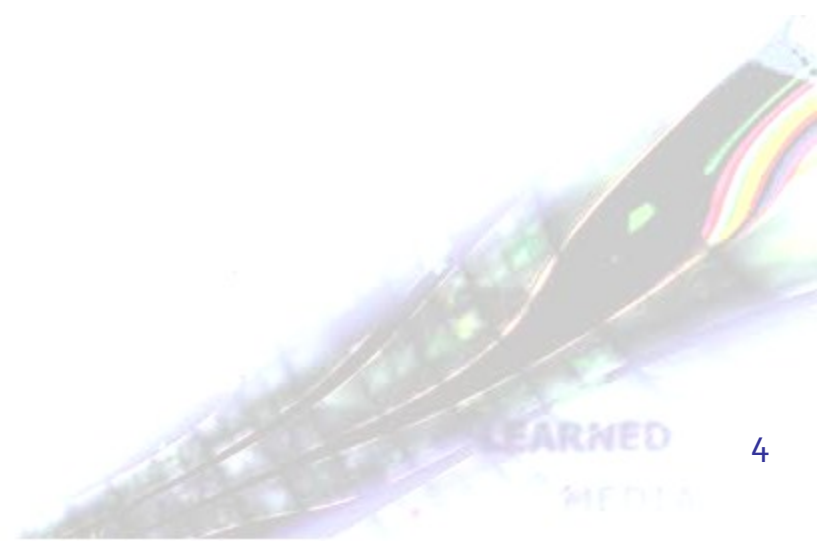
- Data mining

- decision tree induction, entropy, information gain, overfitting, Occam's razor, model pruning, naïve Bayes classifier, KNN, association rules, support, confidence, predictive vs. descriptive DM, numeric prediction, regression tree, model tree, heuristics vs. exhaustive search

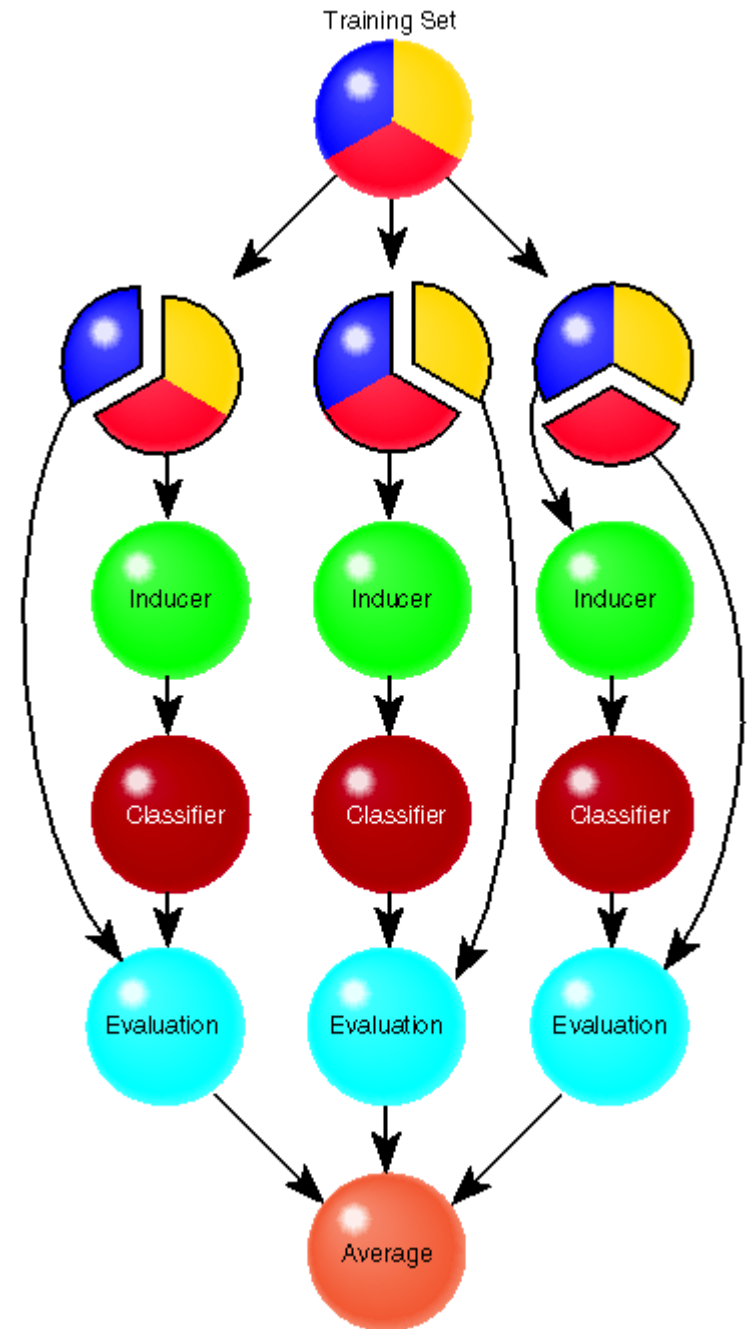
- Evaluation

- Accuracy, confusion matrix, cross validation, ROC space, error, leave-one-out

# Short-sightedness of decision trees



# Cross validation



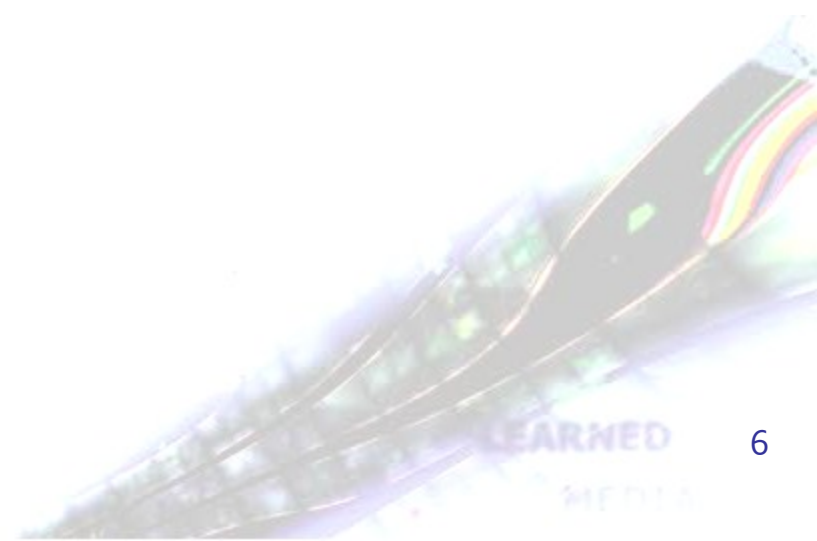
# Predicting with Naïve Bayes

Given

- attribute-value data with nominal target variable

Predict

- the target value of new examples using the Naïve Bayes classifier



# Naïve Bayes classifier

$$P(c | a_1, a_2, \dots, a_n) = P(c) \prod_i \frac{P(c | a_i)}{P(c)}$$

class

value of attribute 1

value of attribute 2

value of attribute n

- Assumption: conditional independence of attributes given the class.

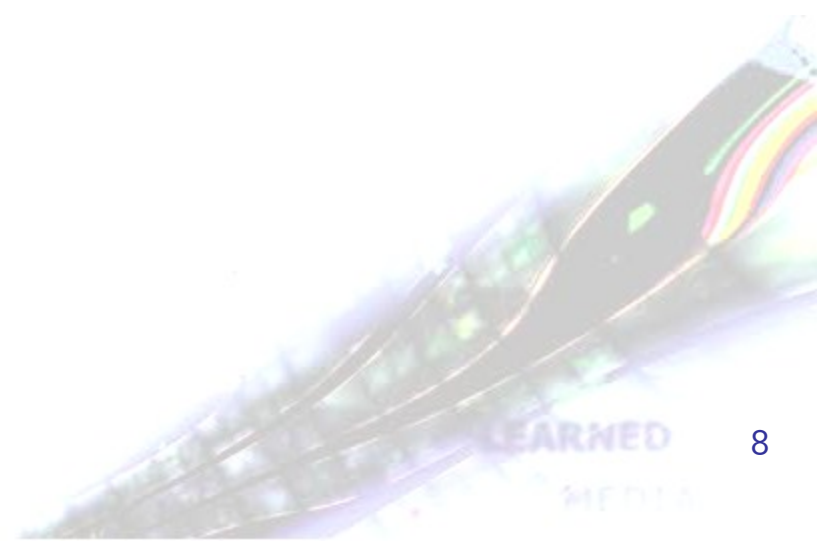
# Naïve Bayes classifier

$$P(c | a_1, a_2, \dots, a_n) = P(c) \prod_i \frac{P(c | a_i)}{P(c)}$$

Will the spider catch these two ants?

- Color = white, Time = night
- Color = black, Size = large, Time = day

Color	Size	Time	Caught
black	large	day	YES
white	small	night	YES
black	small	day	YES
red	large	night	NO
black	large	night	NO
white	large	night	NO





# Naïve Bayes classifier -example

Color	Size	Time	Caught
black	large	day	YES
white	small	night	YES
black	small	day	YES
red	large	night	NO
black	large	night	NO
white	large	night	NO

$v_1 = \text{"Color = white"}$

$v_2 = \text{"Time = night"}$

$c_1 = YES$

$c_2 = NO$

$$p(c_1|v_1, v_2) = p(\text{Caught} = YES | \text{Color} = \text{white}, \text{Time} = \text{night}) =$$
$$p(\text{Caught} = YES) * \frac{p(\text{Caught} = YES | \text{Color} = \text{white})}{p(\text{Caught} = YES)} * \frac{p(\text{Caught} = YES | \text{Time} = \text{night})}{p(\text{Caught} = YES)} =$$
$$\frac{1}{2} * \frac{1}{2} * \frac{1}{4} = \frac{1}{4}$$

# Naïve Bayes - discussion

- What methods can be used for estimating the quality of naïve Bayes predictions?
- How comes that
  - $P(C|a_1, a_2) + P(\text{not } C|a_1, a_2) \neq 1$
- Compare the naïve Bayes classifier and decision trees regarding
  - the handling of missing values
  - numeric attributes
  - interpretability of the model



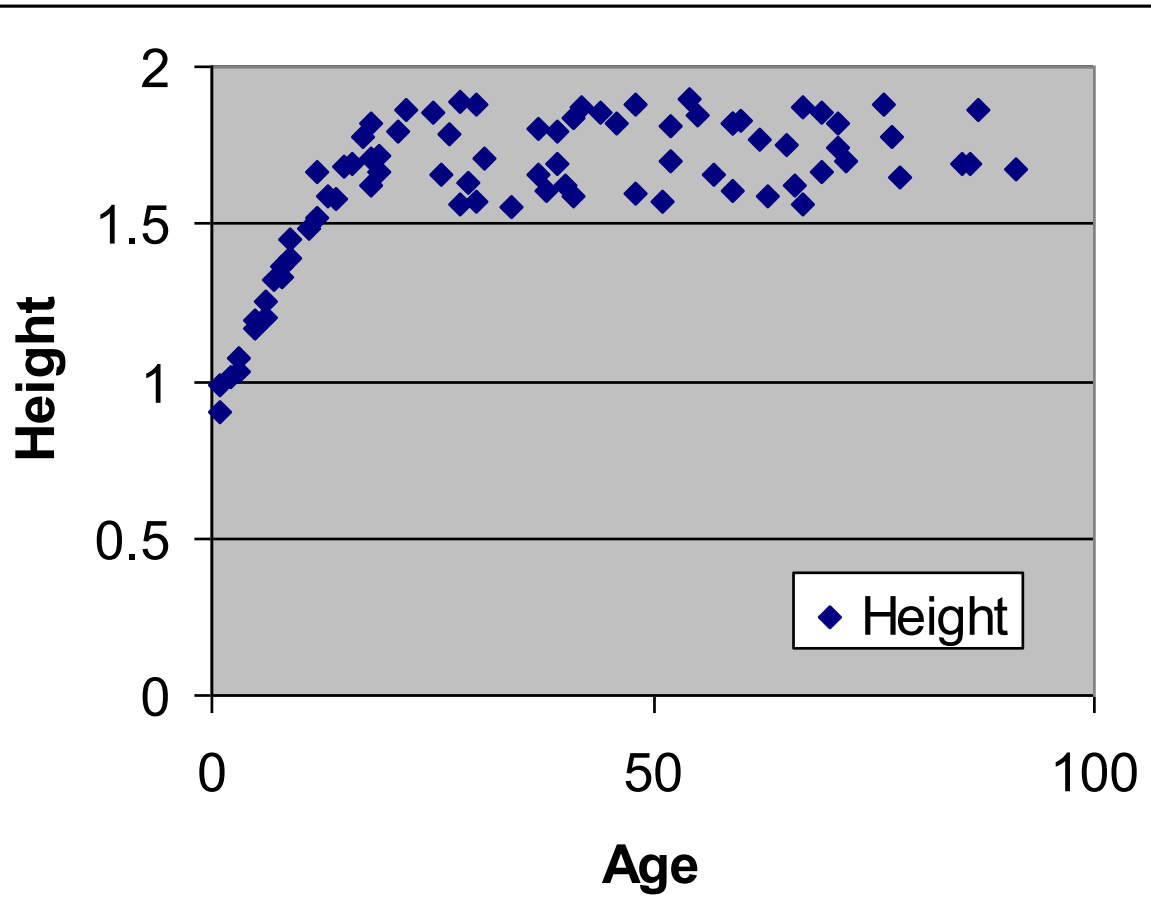
# Numeric prediction

Baseline,  
Linear Regression,  
Regression tree,  
Model Tree,  
KNN

<b>Numeric prediction</b>	<b>Classification</b>
<b>Data:</b> attribute-value description	
<b>Target variable:</b> Continuous	<b>Target variable:</b> Categorical (nominal)
<b>Evaluation:</b> cross validation, separate test set, ...	
<b>Error:</b> MSE, MAE, RMSE, ...	<b>Error:</b> 1-accuracy
<b>Algorithms:</b> Linear regression, regression trees,...	<b>Algorithms:</b> Decision trees, Naïve Bayes, ...
<b>Baseline predictor:</b> Mean of the target variable	<b>Baseline predictor:</b> Majority class

# Example

- data about 80 people:  
Age and Height



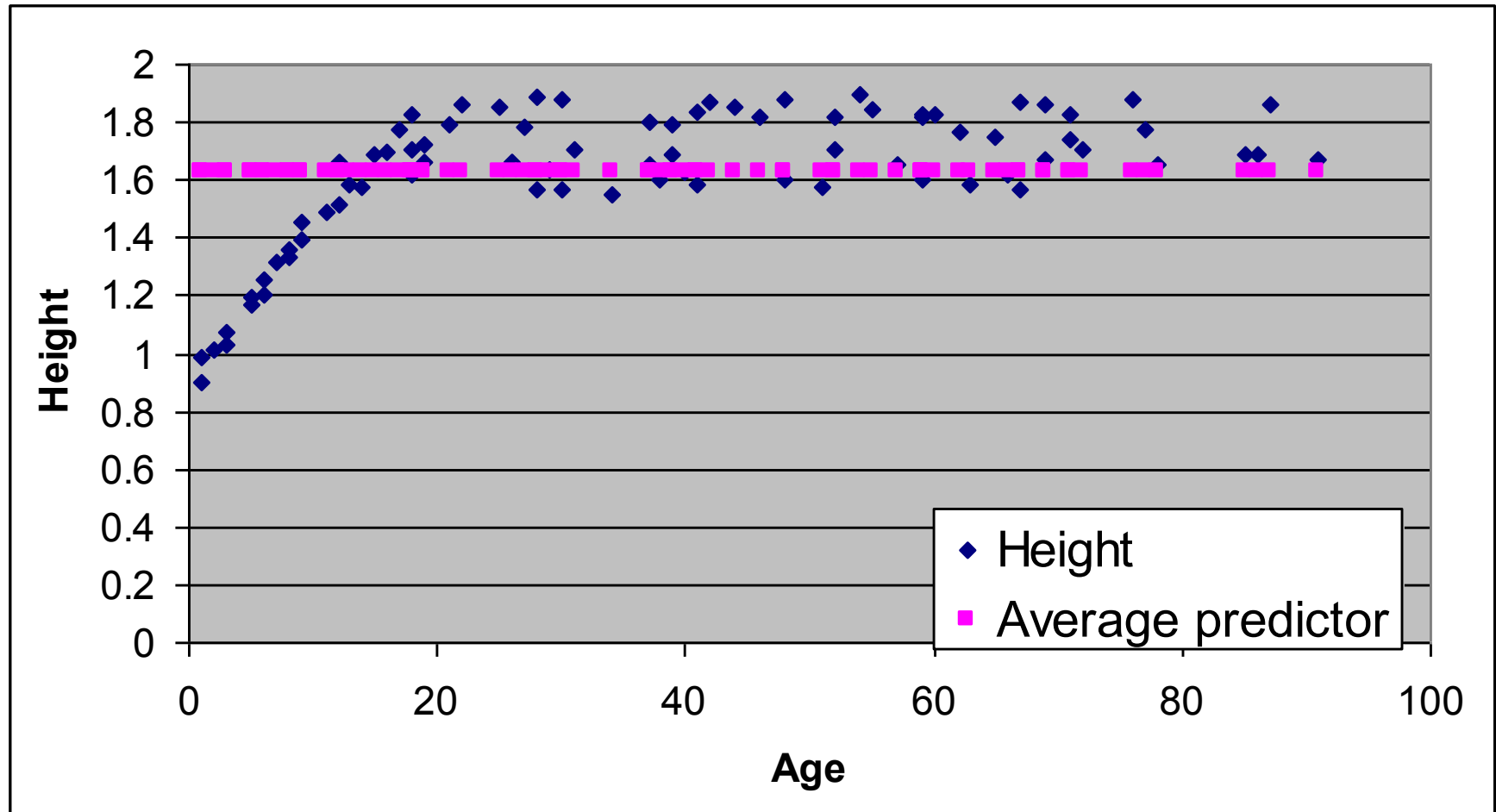
Age	Height
3	1.03
5	1.19
6	1.26
9	1.39
15	1.69
19	1.67
22	1.86
25	1.85
41	1.59
48	1.60
54	1.90
71	1.82
...	...

# Test set

Age	Height
2	0.85
10	1.4
35	1.7
70	1.6

# Baseline numeric predictor

- Average of the target variable



# Baseline predictor: prediction

Average of the target variable is 1.63

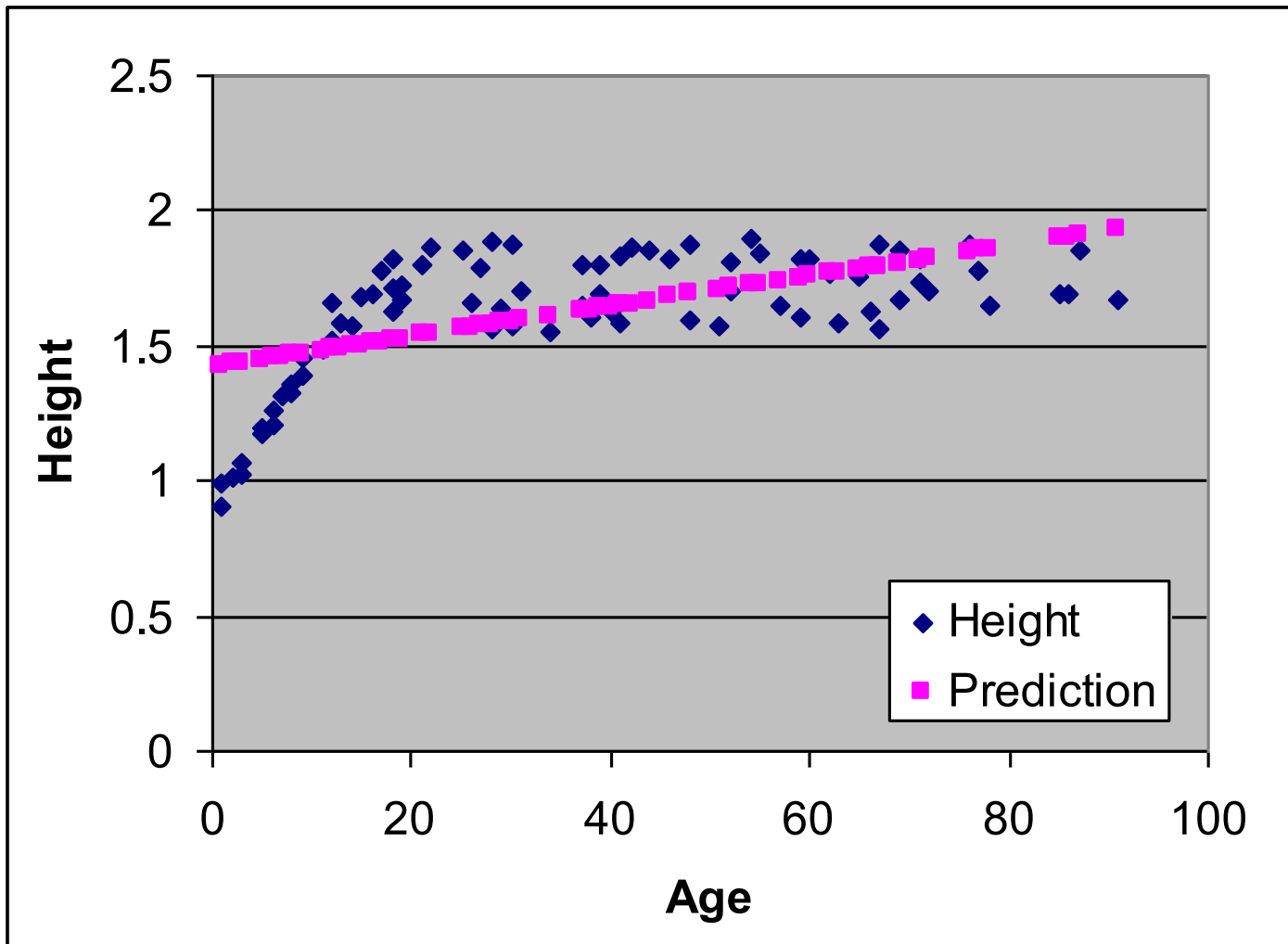
Age	Height	Baseline
2	0.85	
10	1.4	
35	1.7	
70	1.6	





# Linear Regression Model

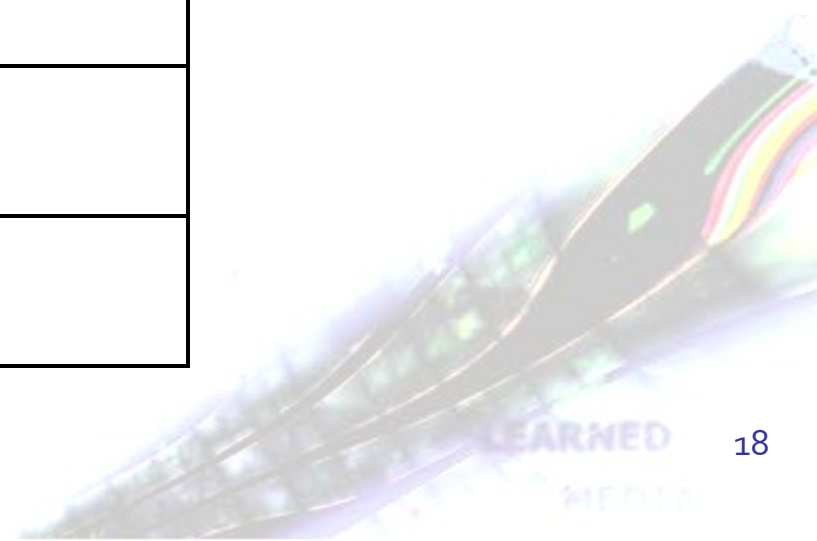
$$\text{Height} = 0.0056 * \text{Age} + 1.4181$$



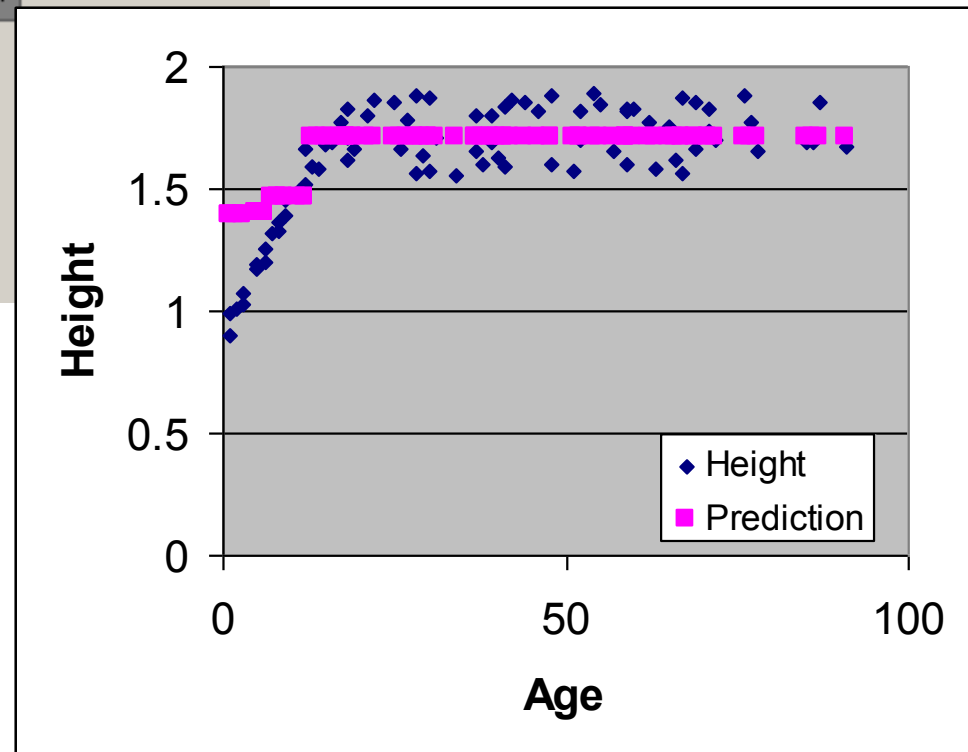
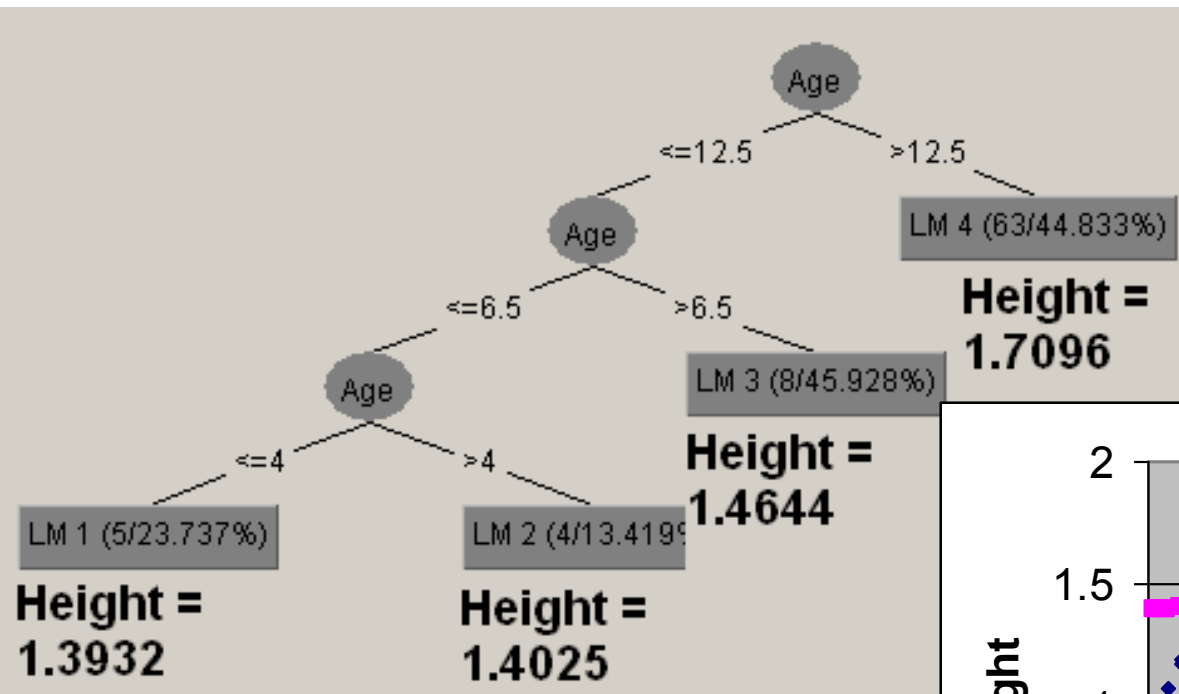
# Linear Regression: prediction

$$\text{Height} = 0.0056 * \text{Age} + 1.4181$$

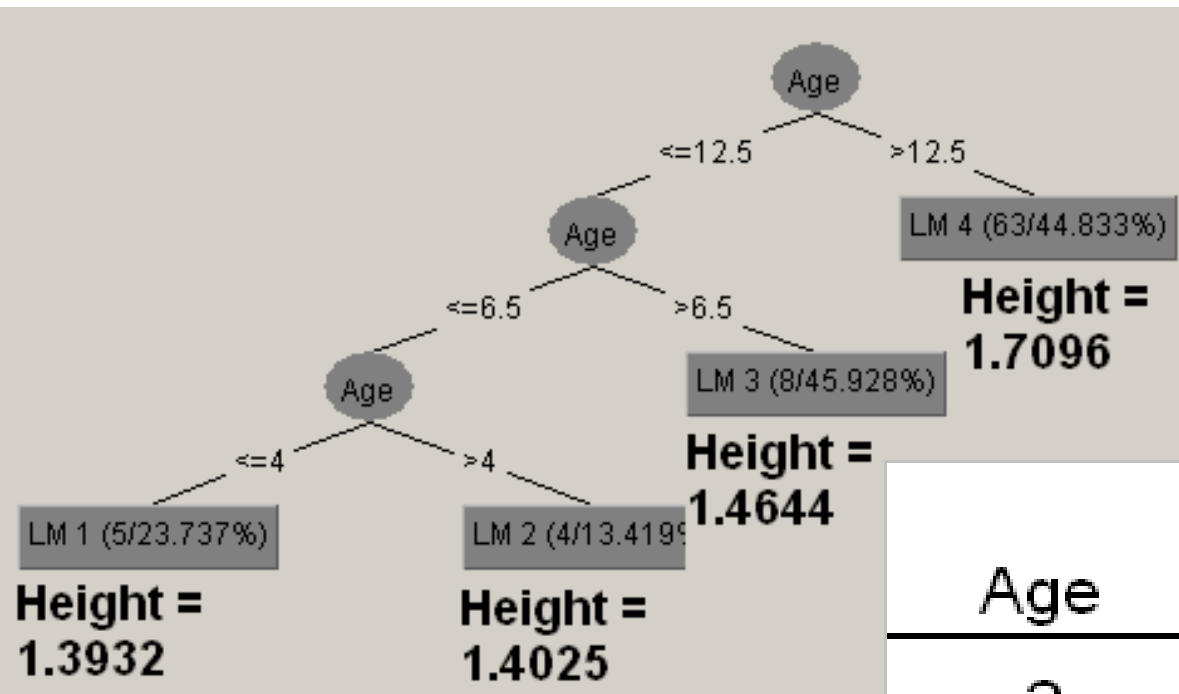
Age	Height	Linear regression
2	0.85	
10	1.4	
35	1.7	
70	1.6	



# Regression tree

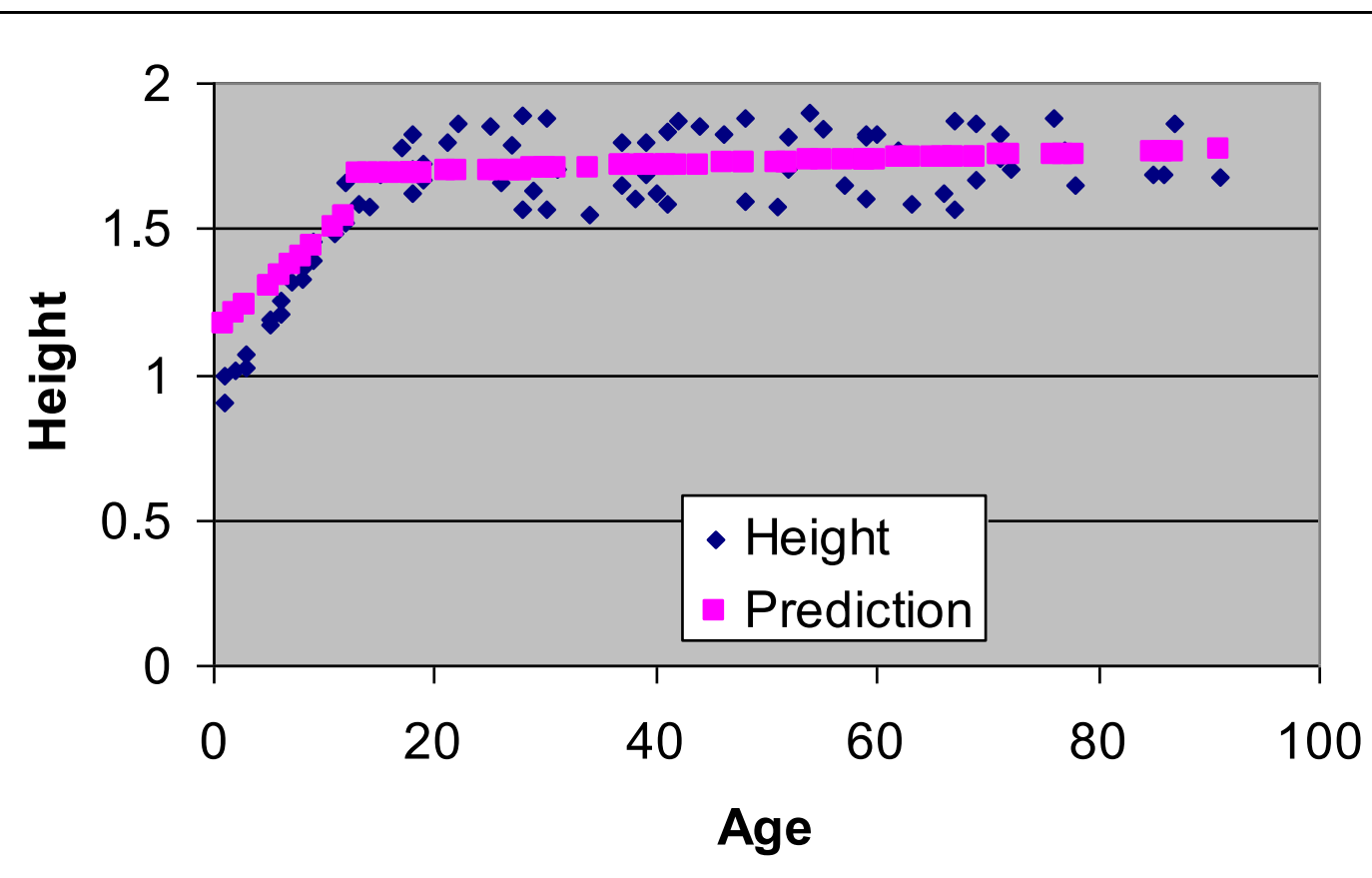
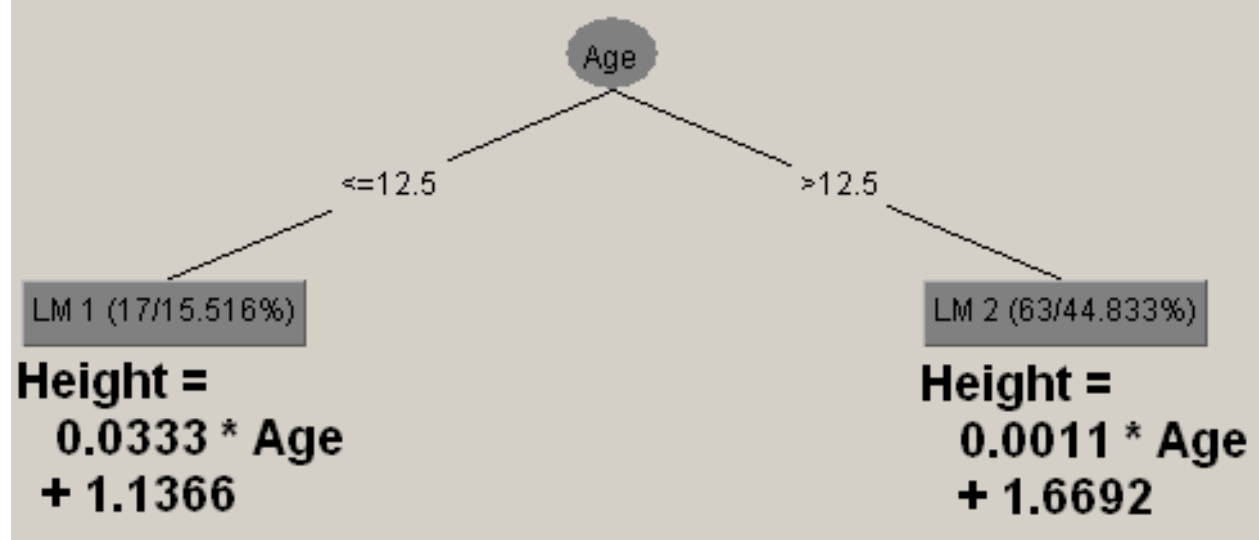


# Regression tree: prediction



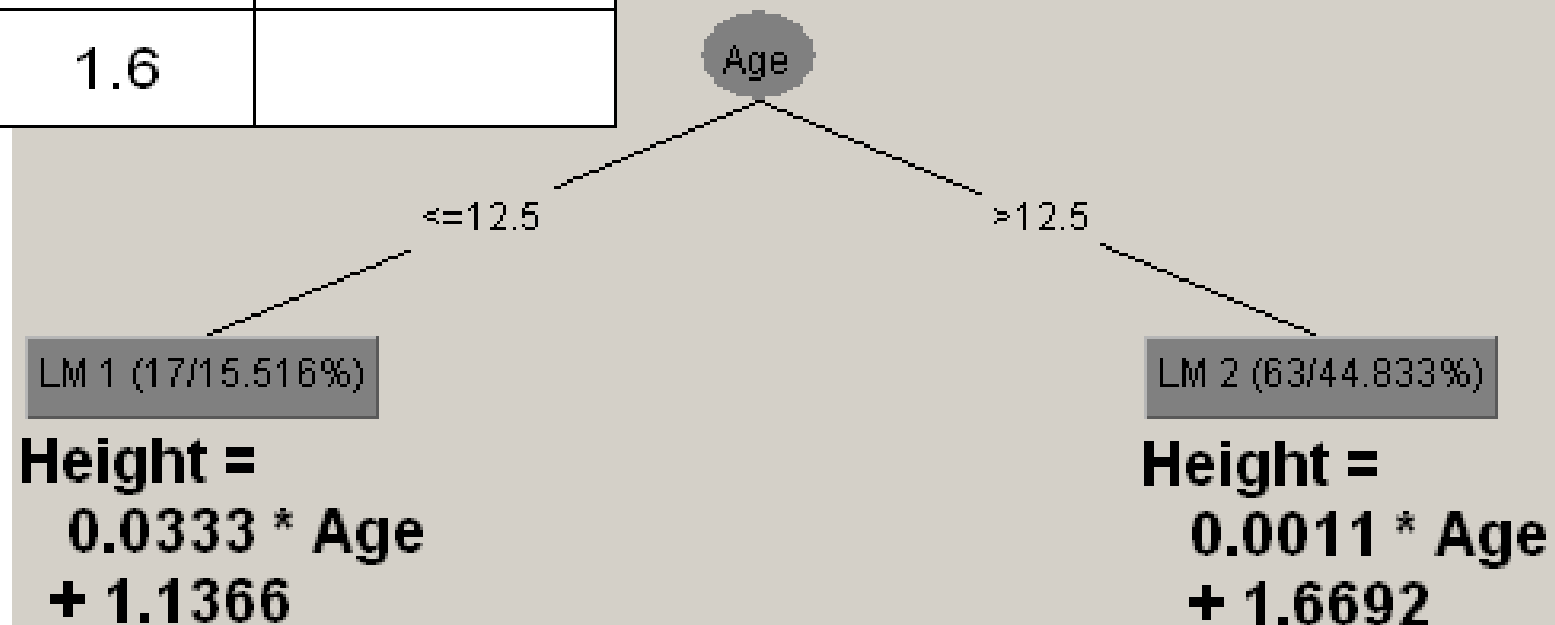
Age	Height	Regression tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	

# Model tree



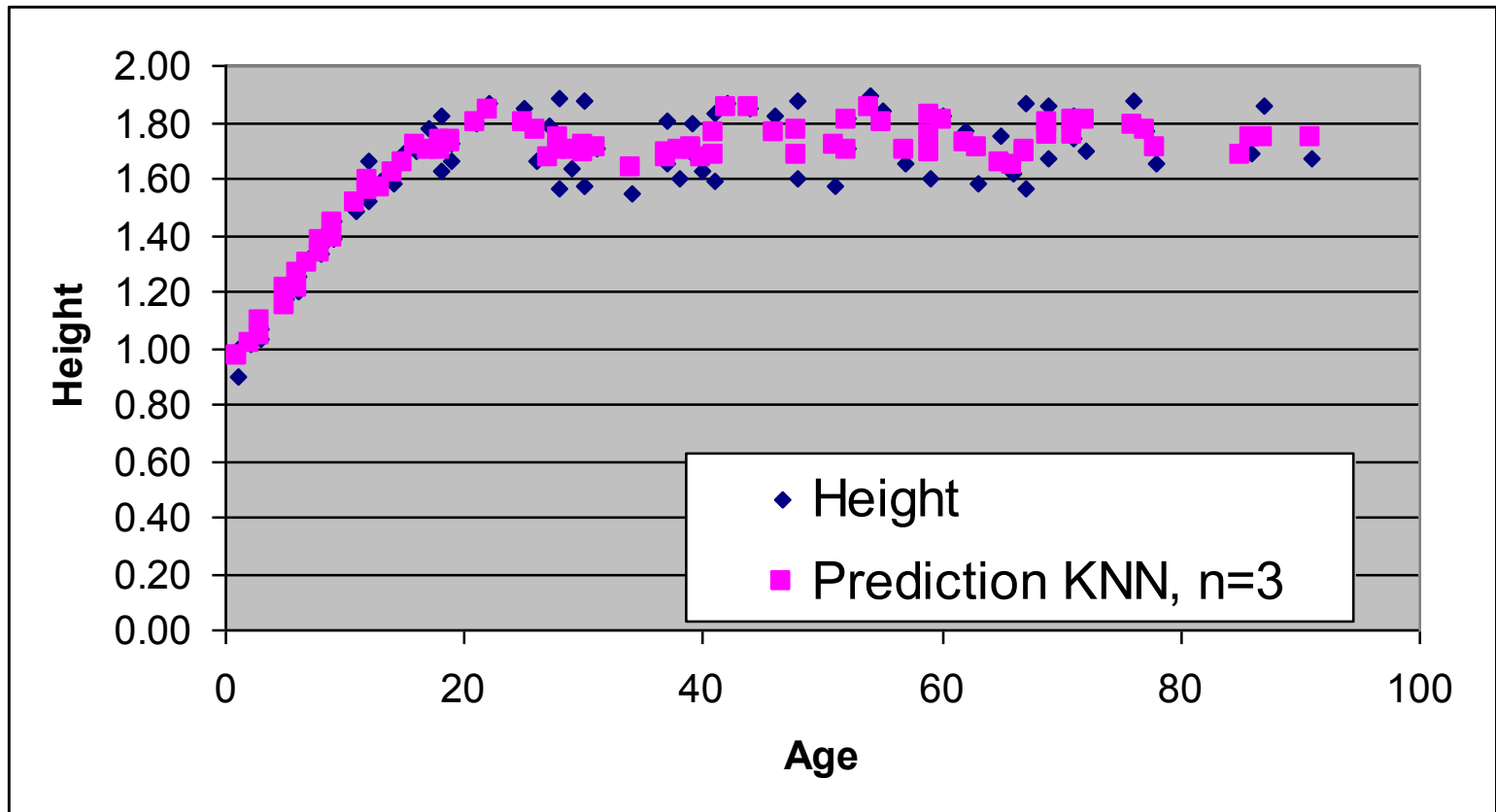
# Model tree: prediction

Age	Height	Model tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	



# KNN – K nearest neighbors

- Looks at K closest examples (by non-target attributes) and predicts the average of their target variable
- In this example,  $K=3$



# KNN prediction

Age	Height
1	0.90
1	0.99
2	1.01
3	1.03
3	1.07
5	1.19
5	1.17

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	



# KNN prediction

Age	Height
8	1.36
8	1.33
9	1.45
9	1.39
11	1.49
12	1.66
12	1.52
13	1.59
14	1.58

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

# KNN prediction

Age	Height
30	1.57
30	1.88
31	1.71
34	1.55
37	1.65
37	1.80
38	1.60
39	1.69
39	1.80

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

# KNN prediction

Age	Height
67	1.56
67	1.87
69	1.67
69	1.86
71	1.74
71	1.82
72	1.70
76	1.88

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

# Which predictor is the best?

Age	Height	Baseline	Linear regression	Regression tree	Model tree	kNN
2	0.85	1.63	1.43	1.39	1.20	1.00
10	1.4	1.63	1.47	1.46	1.47	1.44
35	1.7	1.63	1.61	1.71	1.71	1.67
70	1.6	1.63	1.81	1.71	1.75	1.77

# Evaluating numeric prediction

Performance measure	Formula
mean-squared error	$\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$
mean absolute error	$\frac{ p_1 - a_1  + \dots +  p_n - a_n }{n}$
relative squared error	$\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}, \text{ where } \bar{a} = \frac{1}{n} \sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}}$
relative absolute error	$\frac{ p_1 - a_1  + \dots +  p_n - a_n }{ a_1 - \bar{a}  + \dots +  a_n - \bar{a} }$
correlation coefficient	$\frac{S_{PA}}{\sqrt{S_p S_A}}, \text{ where } S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1},$ $S_p = \frac{\sum_i (p_i - \bar{p})^2}{n-1}, \text{ and } S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$

# Discussion

- List evaluation methods for classification.
- Describe cross validation.
- Compare cross validation, leave-one-out and testing on a separate test set.
- Compare the naïve Bayes classifier and decision trees regarding
  - the handling of missing values
  - numeric attributes
  - interpretability of the model
- How would you compute the information gain for a numeric attribute?
- Can KNN be used for classification?
- How do we avoid overfitting in KNN.
- What do KNN and naïve Bayes have in common?
- Compare numeric prediction and classification.
- Compare decision and regression trees.